Direct and derivative moral responsibility: An overlooked distinction in experimental philosophy

Pascale Willemsen

(Department of Philosophy, University of Zurich, Zuerichbergstrasse 43, CH-8044 Zurich
Pascale.Willemsen@uzh.ch)

Preprint: to appear in Advances of Experimental Philosophy of Free Will

Abstract

Moral philosophers draw an important distinction between two kinds of moral responsibility. An agent can be directly morally responsible, or they can be derivatively morally responsible. Many scholars in the debate believe that direct moral responsibility for an action presupposes that the agent could have acted other than she actually did. However, in some situations, we hold agents responsible even though they could not have acted differently, such as when they recklessly cause an accident or do not take adequate precautions to avoid harmful consequences. Moral philosophers often argue that what we ascribe in these cases is derivative moral responsibility for the action, which results from direct moral responsibility for some other, earlier action. In this paper, I apply this conceptual distinction to the experimental debate about so-called folk-compatibilism or, more precisely, to the question of whether the folk reject the Principle of Alternative Possibilities. I argue that experimental philosophers have failed to consider this distinction when designing experiments and interpreting their results. With the help of three experiments, I demonstrate that intuitions which seem to conflict with the Principle of Alternative Possibilities are best explained by the attribution of derivative moral responsibility. For this reason, these studies do not speak in favour of compatibilism.

Keywords:

Derivative moral responsibility; direct moral responsibility; blame; Principle of Alternative Possibilities; compatibilism

1 Introduction

Consider the following scenario: Jim is a lorry driver on his way home from a long drive. He has been on the road for hours, with two hours still to go. Jim realises that he is getting more and more tired and that it would be wise to take a break. However, the thought of getting home as soon as possible and finishing the day in his own bed is too tempting. Jim keeps driving. A few minutes later, Jim falls asleep for just a few seconds. When he wakes up, he notices that he is only meters away from a car parked next to the road. Even though he hits the brakes immediately and as hard as he can, it is impossible for Jim to avoid the collision. Jim hits the car, and both the car and his lorry are severely damaged. *Is Jim morally responsible for the damage?*

If your intuitions are anything like mine, you believe that Jim is morally responsible for the damage, deserving of blame and punishment, and liable for compensations. But why do we believe that Jim is morally responsible? What is it by virtue of which Jim deserves blame and punishment? Note that Jim never intended to cause the damage. The accident was just that – an accident. Additionally, Jim was unable to avoid causing it. Since the braking distance was longer than the distance to the car when he woke up, it was physically impossible to avoid the accident. So why would we hold him responsible for something he could not have avoided?[1]

The case of Jim exemplifies a situation in which an agent is considered morally responsible for something which, at the time of the action, he could not have avoided. Cases in which an agent is deemed morally responsible despite not being able to act otherwise are used to argue against the Principle of Alternative Possibilities (PAP). Experimental philosophers have examined whether the folk make moral judgments in accordance with PAP. Arguing that philosophical thought experiments cannot suffice to substantiate the adequacy of PAP, they have conducted experimental studies which are inspired by those philosophical thought experiments and aim to provide additional, empirical, and more systematic evidence on whether moral responsibility is dependent on the agent's ability to act otherwise. Among others, John Turri (2017a) argued that the folk seem to be 'natural compatibilists'[2]. In six

[1] We might come closer to the answer if we consider a different scenario in which the agent did not intend the outcome and is unable to avoid causing it. Suppose that John is also a lorry driver on his way home. Despite no known history of any health issues, John suddenly has a heart attack and becomes unconscious. Because he is unconscious, he cannot bring his lorry to a stop. In such a case, I believe that most people do not have the intuition that John is morally responsible for the damage. In the rest of this paper, we will discuss the important differences between Jim and John.

[2] The jump from a violation of PAP to compatibilism is much more complicated than my formulations suggest. For the time being, let us assume that moral intuitions in violation of PAP suggest a compatibilist stance.

original experiments, Turri demonstrated that laypeople are willing to ascribe moral responsibility and blame to an agent who could not have acted otherwise. His empirical results also supported previous experimental findings (e.g. Buckwalter, 2017; Miller & Feltz, 2011; Murray & Lombrozo, 2017; Willemsen, 2018, 2020; Woolfolk et al., 2006).

In this paper, I will take three steps towards demonstrating that this experimental evidence should be considered with caution. First, I will introduce a conceptual distinction between two kinds of moral responsibility, namely *direct* and *derivative moral responsibility* – a distinction which I believe has been overlooked by experimental philosophers. Derivative moral responsibility refers to an agent's moral responsibility for an action or the outcome of an action by virtue of something else they did.[3] Non-derivative, or 'direct' (as it is alternatively termed), moral responsibility denotes an agent's moral responsibility without any intermediate, responsibility-transmitting element. Second, I argue that the stories used in John Turri's original studies allow for the attribution of both direct and derivative moral responsibility. While the attribution of direct moral responsibility would indeed demonstrate that the folk reject PAP, the attribution of derivative moral responsibility would not allow for this conclusion and is compatible with both the acceptance and rejection of PAP. However, which kind of moral responsibility is actually ascribed is unclear. Third, I conducted three experiments to demonstrate that my reservations are not simply theoretical possibilities. Participants' judgments only seem to violate PAP as long as the attribution of derivative moral responsibility is an option. If derivative moral responsibility is less likely to be ascribed, the results no longer support the compatibilist conclusion Turri wished to draw.

I close with the audacious and troublesome claim that much of the experimental evidence to date also fails to draw this conceptual distinction and, more critically, to control for the possibility that laypeople's seemingly compatibilist intuitions are in fact moral judgments about derivative moral responsibility.

2 The Principle of Alternative Possibilities, Direct and Derivative Moral Responsibility

Philosophers typically believe two conditions to be necessary and only jointly sufficient for moral responsibility (Rudy-Hiller, 2018). First, the agent needs some sort of control over what

[3] In the philosophical literature, derivative moral responsibility is discussed as the result of a tracing strategy. This strategy play a major role in many theories of responsibility (see, e.g Khoury 2012, King 2014, Shabo 2015, and Timpe (2011).

they are doing – the *control condition* of moral responsibility. Following Frankfurt (1969), one popular way to spell out this control condition is the Principle of Alternative Possibilities:

> (PAP): An agent is morally responsible for what she has done only if she could have done otherwise.

But what does it mean to be responsible for something *one has done*, and what does it mean that an agent *could have done otherwise* (see Miller & Feltz, 2010 for a similar discussion)? According to one understanding of PAP, moral responsibility requires that an agent's action result from her own choice among a variety of options. Consequently, an agent is morally responsible for the action she chose if there were alternative courses of actions the agent could have chosen instead.[4] Note that this understanding focuses on the agent's *action* and the *situational circumstances when initiating the action* – the *Principle of Alternative Actions* (see Willemsen, 2020, for a discussion). A different understanding of PAP does not focus on the circumstances under which the action was initiated, but rather takes the action to be defined by its *consequences* (for such an understanding of PAP, see, among others: van Inwagen, 1983, 1999; Sartorio, 2005). An agent is morally responsible for *killing* a man, for example, if the consequence of her action is the death of a person, and if this death could have been prevented. If the victim would have died no matter what, the agent is not morally responsible for the death. This is the *Principle of Alternative Outcomes* or, as Miller and Feltz (2011) termed it, the *Principle of Possible Prevention*.[5]

A second necessary condition for moral responsibility is the *epistemic condition*. An agent requires some relevant sort of awareness of what they are doing. Suppose that Tom pushes a button in his new office which he believes will turn on the light. In fact, the button administers severe electric shocks to a person in another room – something that Tom could not have possibly known. Even though Tom has full control over pushing the button, he is not sufficiently aware of what he is doing by pushing it to qualify as morally responsible. Relatedly,

---

[4] Please note that this clarification is still not sufficiently sharp. For instance, according to an unconditional reading, "could have chosen instead" means that the agent could have chosen otherwise even if all antecedent conditions had been the same. Defenders of a conditional reading, usually compatibilists, understand "chould have chosen otherwise" as saying "if the something leading to the decision had been different". In a recent paper, Huber et al. (forthcoming) provide empirical evidence that this distinction matters for research on folk intuitions.

[5] Typically, having alternative courses of action available goes hand-in-hand with being able to bring about alternative outcomes. Choosing a different course of action usually leads to different outcomes. However, it is possible for an outcome to be determined in a way that, no matter what an agent does, the same outcome will occur. Suppose that a patient is very ill and suffers from an incurable disease. No matter what the physicians do, the patient will die. Here, the physicians can act differently without being able to bring about alternative outcomes. Alternatively, think of a case in which a person can only act she actually does, but whether or not she brings about a certain outcome largely depends on other factors beyond her control. Such cases are critical to discussions about the moral significance of luck.

we typically consider young children blameless for hurting others, as we believe they lack sufficient understanding of the moral relevance of their actions.[6]

Only if both the control condition and the epistemic condition are met can the agent be morally responsible for their actions. This kind of responsibility is often referred to as *ultimate*, *true,* or *direct moral responsibility* (e.g. Levy, 2017; Matheson, 2019; Mele, 2020; Strawson, 1994). In the following, I will use the term '*direct moral responsibility*'.

Consider the example from the beginning of this paper: Jim is tired, continues driving, falls asleep, and wakes up only to realise that hitting a parked car is unavoidable. Is Jim morally responsible for the accident and the damage he caused? On the face of it, Jim does not fulfil the control condition at the time of the accident. When he wakes up, only a few meters away from the parked car, Jim cannot decide to bring his own car to an earlier stop and thus not to hit the parked car. Nevertheless, it is intuitively plausible that Jim *is* morally responsible. What should we make of this? First, it might be argued that our moral intuitions clearly suggest that we implicitly reject PAP as incorrect and that a lack of alternative possibilities is considered compatible with moral responsibility. Alternatively, one could maintain the (conceptual or metaphysical) truth of PAP by discarding our intuition as somehow flawed or biased.

Here is a third way to explain our intuitions: Philosophers such as Zimmerman (1997), Pereboom (2012), Rosen (2003), and others have distinguished between direct moral responsibility and what they call 'derivative' moral responsibility[7]. An agent is *derivatively* morally responsible for X if they are considered morally responsible for X at t_0 *in virtue of* being (directly) morally responsible for Y at t_{-1}. While one might agree that Jim cannot be *directly* morally responsible for the accident, he can be *derivatively morally responsible* for the accident in virtue of being directly responsible for something else, namely for driving in an impaired state. Appropriate responses to his tiredness would have been to take a break, get some fresh air, take a short nap, and then continue the drive. The reason we feel that Jim is

[6] Both the control and the epistemic conditions are subject to intense philosophical debate, and various specifications of them have been offered. It is beyond the scope of this paper to do justice to this debate. For an overview, see Rudy-Hiller, 2018.

[7] The distinction also plays a key role in explaining how an agent can be morally responsible for unwitting omissions (e.g., Nelkin & Rickless, 2017; Rosen, 2003, 2004; Sartorio, 2007). Suppose that I promise to buy groceries on my way home from work. When I come to the crossing at which I am supposed to turn left to the supermarket, I take a right turn as I would otherwise usually do. I forget to go to the supermarket. In this case, it seems that I was not aware of what I was doing – I was not aware that I was breaking my promise to buy groceries. Yet, it seems reasonable to hold me responsible for the missing groceries. The moral responsibility here ascribed is, again, derivative and can be traced back to an earlier point in time at which I fulfilled all necessary conditions for moral responsibility. Knowing how forgetful I am, I should have made myself a reminder or paid more attention to my duties (see e.g., King, 2009 and Robichaud & Wieland, 2017 for critical positions; see Rudy-Hiller, 2018 for an overview).

morally responsible for the accident is that we ascribe responsibility for something over which, in our estimation, he had control. Jim is directly morally responsible for driving even though he was too tired, and thus he is derivatively morally responsible for the accident and the damage he caused by virtue of being directly responsible for driving in his impaired state.

Suppose that this explanation is a psychologically adequate description of why Jim is considered morally responsible. Can we conclude that by judging Jim derivatively morally responsible for the accident, we reject PAP? Not at all. A violation of PAP requires an agent to be considered directly morally responsible for X at t_0 and also unable to avoid performing X at t_0. However, if my premise is correct, Jim is derivatively morally responsible for X at t_0 in virtue of what he did at t_{-1}, a time during which we have no reason to assume that Jim was not in control of his actions. Thus, at no point are our intuitions in conflict with PAP.

Distinguishing direct and derivate moral responsibility can explain why we sometimes blame an agent even though the necessary conditions for (direct) moral responsibility are violated. It may further help us to be more specific as to why or by virtue of what we consider an agent responsible, and to better understand where the wrongness of an agent's behaviour lies. In the following, I argue that this conceptual clarity is indispensable, especially when designing experimental studies on folk morality and interpreting their results.

3 Experimental Findings: Direct or Derivative Moral Responsibility?

While moral responsibility is a key topic of traditional moral philosophy, it has also attracted the attention of experimental philosophers. The control condition in particular has been subject to experimental studies because of its role in compatibilism of moral responsibility and free will, as well as related issues such as moral luck and the so-called ought-implies-can principle.

When investigating the folk's intuitions, the usual strategy is to present participants with an experimental stimulus in which an agent cannot do other than they actually do and, therefore, the control condition is violated. Quite often, these experimental stimuli are adapted from a philosophical thought experiment, such as versions of Frankfurt cases (Miller & Feltz, 2011; Murray & Lombrozo, 2017; Nahmias et al., 2005; Turri, 2017b; Willemsen, 2020). Participants read the stimulus and are subsequently asked whether the agent is morally responsible or to blame. If participants agree that the agent is responsible despite their inability to act otherwise, this is taken as support for folk compatibilism[8].

[8] Please note that this interpretation is in fact inadequate. Compatibilism is a thesis about the compatibility of moral responsibility or free will with determinism. While some philosophers believe that determinism is the most

In the following, I focus on a recent paper by John Turri (2017a) entitled 'Compatibilism can be natural'. Turri presented participants with one of the following two stories:

> **(Evaluation)** A woman is evaluating her employee's performance. The employee performed excellently. Given the current condition of the woman's brain, it is physically impossible that she can give the employee a positive evaluation. As a matter of brain chemistry, it is literally impossible that she can give the employee a positive evaluation. She will give the employee a negative evaluation.
>
> **(Delivery)** A man promised to deliver a package by 4pm. He just got on the freeway. Given current traffic conditions, it is physically impossible that he can deliver the package by 4pm. As a matter of physics, it is literally impossible that he can make it by 4pm. He will arrive late.

In both stories, the agent is described as violating the control condition. The woman is no longer in control over whether she gives her employee a positive or negative evaluation due to the current condition of her brain; the delivery man is not in control over the time at which he delivers the parcel, as he got stuck in traffic. To lend support to the claim that the folk reject PAP, Turri needs to show two things:

1. Participants say that the agent could not have acted other than performing X at t.
2. Those participants who believe that the agent could not have acted otherwise still ascribe moral responsibility for X at t.

Turri (2017a, p. 79) does find this evidence across a variety of experiments, and concludes:

> The present experiments provide the best evidence to date for natural compatibilism, completely avoiding weaknesses of prior work on the topic […] I used brief, plain, tightly matched, and anodyne stimuli, tested multiple narrative contexts, and included multiple measures to assess how participants understood key variables. Participants understood the stimuli in the relevant way. The manipulations were credible and effective.

Turri is correct that he managed to avoid many of the methodological shortcomings that have plagued other studies.[9] Given the superiority of his approach, his paper does have the potential to provide excellent evidence in favour of the position that the folk reject PAP. However, at least one further condition must be added:

3. The moral responsibility that participants ascribe is *direct* moral responsibility.

---

systematic violation of the ability to act otherwise, not all do. Many compatibilists argue that the ability to do otherwise is a necessary precondition for moral responsibility (and, thus, that PAP is true); they also argue that determinism is compatible with having this ability. Thus, they are compatibilists, and maintain the truth of PAP. In the following, I will try to avoid such misleading formulations. I believe that Turri's paper is much more adequately framed as addressing the equally-interesting question of whether the folk reject PAP.

[9] It is beyond the scope of this paper to discuss these methodological flaws in detail. To name a few, some papers were criticised for having used experimental stimuli and descriptions of causal determinism that laypeople did not fully understand in the relevant way. Other studies failed to test whether participants actually believed that the agent could not have done otherwise. In yet other studies, vignettes were confounded and not sufficiently tightly matched (see Turri, 2017a for a more detailed discussion).

To provide any evidence that can speak to whether the folk accept or reject PAP, we need to make sure that when participants hold the agent morally responsible, they ascribe direct moral responsibility for the action that is described as unavoidable. Why? To tests whether the folk reject PAP, we need a situation in which the control condition is violated, and the agent's behaviour was without alternatives. Only if participants ascribe moral responsibility in violation of the control condition is there evidence that PAP is rejected. However, if participants ascribe derivative moral responsibility instead, they have shifted their attention away from the situation that violates the control condition. Therefore, we cannot make any inferences as to whether PAP is rejected, because PAP is no longer under investigation.

The demand I wish to make here is a methodological one, requiring us as experimental researchers to clearly show that we have tested the relevant moral intuition. While thorough research and the exclusion of all potential confounds is certainly what we do and should aim for, one might wonder whether I exaggerate a minor fluke that is unlikely to cause any actual trouble. Should we really expect that in a design as simple as Turri's, people misunderstood what they were supposed to do?

Despite the importance of the distinction between direct and derivative moral responsibility in the philosophical literature, there is no natural way to express the two different kinds of moral responsibilities in ordinary language. We typically ascribe blame to agents, hold them responsible, and punish or condemn them for causing harm to others. Usually, we do not specify whether this moral responsibility is direct or derivative in nature. This lack of conceptual clarity can cause problems in two ways. Firstly, a lack of discriminating vocabulary might cause the participants confusion as to what kind of moral responsibility they are being asked to ascribe. Participants might assume that there are various things for which we could blame the agent, all of which are somehow connected to the outcome. For some things the agent is directly responsible, while for others the agent is responsible only by virtue of something else. However, participants might wonder on what basis moral responsibility should be ascribed, perhaps in the following sense:

1. The agent could not have acted other than they actually did when they performed X at t_0.
2. Hence, the agent is not morally responsible for X at t_0, as they could not have acted otherwise (*a judgment in line with PAP*).
3. However, the agent performed some other blameworthy action Y at t_{-1}that led them to perform X at t_0.

4. Therefore, the agent is morally responsible for X at t_0 *by virtue of* being responsible for Y at t_{-1}, despite not being able to do otherwise at t_0 (*a judgment seemingly in violation of PAP*).[10] In an experimental setting, such a line of reasoning will lead to judgments that seem as if participants reject PAP and attribute moral responsibility despite a recognition of a lack of alternative possibilities. However, since the moral responsibility judgment is a derivative one, it is not incompatible with PAP at all.

Alternatively, participants might not possess this level of self-reflection, and lack access to the reasons for which they blame others. The fact that there is no discriminatory vocabulary available could be taken as direct evidence that, while *philosophically* relevant, the distinction is irrelevant in our everyday lives. As a consequence, upon reading the test query, participants may have ascribed moral responsibility without wondering (or caring) what it is they ascribed moral responsibility for – they had the intuition that the agent was blameworthy for something, and that is the response they provided.

Let us examine Turri's vignettes to see if they leave room for derivate moral responsibility attribution. In *Evaluation*, a woman is described as unable to act other than to give her employee a negative evaluation as a matter of her brain chemistry. It is unclear whether participants are familiar with the science underlying the test vignette, or are aware of the implications of the current conditions of a person's brain chemistry. If they are not, participants might have tried to make sense of this information by making additional assumptions. For instance, they might have believed that the woman caused her own inhibition by not sleeping enough or by taking drugs which now impair her judgment and proper brain function. Since Turri does not specify why the woman's brain is in this state, the woman might be considered morally blameworthy because she could and should have avoided whatever has caused this brain state – or so participants might have reasoned. The woman is, therefore, considered derivatively morally responsible for not giving the employee a positive evaluation by virtue of causing or allowing her brain to be in this malfunctioning state. Such results, though, do not test PAP, and therefore do not allow any inferences as to whether the folk reject it.

[10] Members of the Lund-Gothenburg Moral Philosophy Group pointed out to me that this explanation, while convincing, might make too charitable an assumption about laypeople's moral cognition. The story I offer here assumes that people go through a rational reasoning process in which they make assumptions about what most probably led to the agent's situation. Alternatively, one would have to think that people are just mean, unreflective blaming machines searching for validation of their outcome-triggered desire to blame.

Similarly, in *Delivery*, a man is described as unable to deliver a parcel by 4 pm due to being stuck in traffic. Turri succeeds in telling a story with which people are familiar, just as he claims is necessary for reliable experimental results. However, people might be *too* familiar with the story. We usually know very well why we are late for an appointment, and it is usually poor planning. Participants might have believed that the man should have foreseen the possibility of a traffic jam, and that he simply left too late. They might also have believed that he should have checked the route in advance, been more alert to the traffic news on the radio, or taken other preparatory measures. Again, if participants enrich the story in this way, it is only reasonable to ascribe derivative moral responsibility. Unfortunately, nothing in the original vignette rules out these (mis)interpretations. Turri's results, thus, would not suggest that the folk reject PAP.[11]

Turri himself seems to have been aware of the possibility that his vignettes could be interpreted in this way. In the general discussion (Turri, 2017a, p. 79), he voices what I take to be a variation of my worry:

> It might be wondered whether people attribute the relevant moral status because they believe that at some point in time, not described in the scenario, the agent could have done something that would have prevented his subsequent inability. If so, the objection continues, none of the results would support natural compatibilism.

Turri does not believe this objection to be powerful, and argues that 'if natural incompatibilism was true, then it seems unlikely that participants would respond as the objection envisions'. I believe this response is mistaken. The worry his critic (in this case, me) has is not that the folk are natural *in*compatibilists or, as I have reframed the problem, that they accept PAP. The worry is that, be the folk as they may, the experimental design is unable to provide evidence in either direction. Nevertheless, I am sympathetic to his concluding remark that if an objection like mine is correct and participants do trace back moral responsibility, then 'ordinary social cognition might never confront the issue of compatibilism or incompatibilism'. I believe this is in fact a possibility worth exploring in future research and should be taken seriously, as it would raise serious questions about whether experimental

[11] Note that these two stories are asymmetric with respect to the two version of PAP. *Evaluation* creates a scenario in which the Principle of Possible Prevention of an unfortunate outcome is at issue. The woman can only give the employee a negative evaluation – neither a positive nor a neutral one. She therefore cannot prevent the outcome. At the same time, it seems that the Principle of Alternative is at issue as well. She cannot act other than to give the employee a negative evaluation, and one doubts that she is in control over this action in a relevant sense. Since she believes that the employee performed excellently, her actions stand in conflict with her mental state. In *Delivery*, such a conflict is not described, and the delivery man seems to be in perfect control over his behaviour. It also seems that his inability to deliver in time is external to him. All of these asymmetries might have an effect on participants' interpretations of the story (see Willemsen, 2020 for a discussion).

studies on folk compatibilism are a worthwhile endeavour. While I cannot provide a satisfactory investigation of social cognition more generally, I assume for the sake of argument that ordinary social cognition at least sometimes confronts issues related to PAP, and I hope to offer some evidence supporting the need for a reservation which Turri discards much too quickly.

To test whether this alternative interpretation can account for Turri's results, I conducted three experiments. In Studies 1 A and B, I tested whether the agent is held morally responsible in a derivative sense. In these studies, I manipulated two things. First, I replicated Turri's original design and added a follow-up question. Participants who indicated that the agent could not act other than they did *and* that the agent is responsible (a judgment in potential violation of PAP) were asked to explain their moral judgment. If my reasoning is correct, participants would explain their judgments through additional assumptions about how the agent could have prevented their own inability. Second, I created an additional test condition with a manipulation of the original vignettes. These manipulated versions provided information that the agent is not (or is less obviously) to blame for causing their inability (see Table 1 for details). If causing an inability to act otherwise is required for people to blame the agents in the original vignettes, describing the agents' inabilities as not self-induced, and therefore blameless, should significantly reduce blame judgments. I tested these manipulations for two different morality queries that can be found in the literature, namely for the agent's blameworthiness for the outcome (Study 1 A) and for the agent being morally responsible (Study 1 B). In deference to the concern that my results in Studies 1 A and B might have been due to the stimuli doubling in length or the introduction of new factors to the stories, I address this possibility in Study 2 by comparing participants' responses to more closely-matched vignettes.

4 Studies 1 A and B

The experimental design for Study 1 built on Turri's original design. The experimental design and all prediction and statistical analyses were pre-registered with the Open Science Framework (*https://osf.io/82ems/?view_only=c229aede0d19439b83bcddb31ad938da*). Since there is no general consensus as to whether 'blame' or 'moral responsibility' provides a more adequate measure for moral responsibility, I tested two responsibility questions. I followed Turri in this decision. In Study 1 A, the responsibility query asked whether the agent is *to blame*; in Study 1 B, the responsibility query asked whether the agent is *morally responsible*. The experiments were identical in all other respects.

The experiment was motivated by the following predictions:

1. I expected to replicate Turri's original results when using the original vignettes.
2. Most participants who give seemingly compatibilist responses will explain their judgment by indicating that there was something the agent could have done to ensure that they would keep their promise or provide an adequate evaluation.
3. For both vignettes, people will blame the agent under No Self-Induced Inability conditions significantly less compared to the original vignettes.
4. For both vignettes, people will still judge the agent under No Self-Induced Inability conditions as unable to perform the action.

Methods for Studies A and B

I utilised a 2 (vignette: Delivery vs. Evaluation) x 2 (condition: Original vs. No Self-Induced Inability) between-subjects design. The Original condition was identical to Turri's original design for both vignettes (Section 4). In the No Self-Induced Inability condition, I added information which made it clear that the agent's inability was not the result of his or her own recklessness (see Table 1).

Table 1: Modified vignettes used in the experiment. Underlined sections represent additions made to the original vignettes.

	Delivery	**Evaluation**
No Self-Induced Inability	A man promised to deliver a package by 4 pm. <u>He planned his route carefully and left very early, so that he would have plenty of time to get to his destination.</u> He just got on the freeway<u>, when two trucks collide before him. The freeway is blocked, and the police inform everyone that it will take at least until late in the night to clear the freeway. Unfortunately, this freeway is the only street that leads to the destination of the package.</u> Given current traffic conditions, it is physically impossible that the man can deliver the package by 4 pm. As a matter of physics, it is literally impossible that he can make it by 4 pm. He will arrive late.	A woman is evaluating her employee's performance. The employee performed excellently <u>and the woman is resolved to and about to give her employee a very good evaluation.</u> <u>Unfortunately and unbeknownst to her, the woman suffered a minor stroke before she began the crucial part of the evaluation. This stroke changed the current condition of the woman's brain.</u> Given the current condition of the woman's brain, it is physically impossible that she can give the employee a positive evaluation. As a matter of brain chemistry, it is literally impossible that she can give the employee a positive evaluation. She will give the employee a negative evaluation.

The questions that I used were identical to Turri's original ones and were answered on a rating scale from 1 ('strongly disagree') to 7 ('strongly agree'):

1. *Ability*: The man could still deliver the package by 4 pm
2. *Responsibility*: The man is to blame (*is morally responsible*) for the time he delivers the package.
3. *Likelihood*: On a scale of 0% to 100%, how likely is it that the man will deliver the package by 4 pm?

When subjects gave an answer to the ability question of 4 or below (indicating indifference or disagreement, respectively), and at the same time gave an answer to the responsibility question of 4 or above (indicating indifference or agreement, respectively), I presented them with the following additional questions (with Time only shown for *Delivery*):

4. *Explanation*: Your judgment indicates that you believe the man is to blame for the time he delivers the package. Please explain your judgment by choosing the option that best expresses your intuition:

   The man is to blame for the time he delivers the package…

   A) … although there was nothing he could have done to ensure he would deliver the package in time.

   B) …because there was something he could have done to ensure he would deliver the package in time.

5. *Time*: Please tell us your best guess of the time the man got on the freeway. Please use the following format: 11:32 pm (hour:minutes am/pm) (do not forget am/pm!)

Study 1 A

Participants

773 participants were recruited through the UK-based internet platform *Prolific* (https://www.prolific.ac). All participants were compensated for their participation (0.25 GBP, estimated 7.50 GBP per hour). All participants were native speakers of English and had not previously participated in an experiment using the same vignettes. I excluded 57 participants from the analysis for either failing the attention check, not completing the survey, or finishing the survey in less than 40 seconds (please see pre-registration for further details). Results are reported for 716 participants (M = 34.37, SD = 12.17, 56% female, 44% male).

Results

Ability and Responsibility Ratings

I conducted t-tests against the midpoint of the scale (4) for ability ratings. Replicating Turri's original results, participants' ability ratings were significantly below the midpoint of the scale for both Original conditions (Delivery: M = 1.14, t = -61.16, p < .001; Evaluation: M = 2.19, t = 14.1, p < .001), indicating that they judged the agent unable to perform their obligation. T-tests against the midpoint revealed that blame ratings did not significantly differ from the scale midpoint in the Original condition (Delivery: M = 3.84, t = -1.18, p = .24; Evaluation: M = 4.08, t = 0.52, p = .06) (see Figure 1).

In contrast, for the No Self-Induced Inability condition, t-test against the midpoint of the scale (4) revealed that both ability (Delivery: $M = 1.35$, $t = -30.28$, $p < .001$; Evaluation: $M = 2.42$, $t = -9.98.1$, $p < .001$) and blame ratings (Delivery: $M = 1.31$, $t = -38.37$, $p < .001$; Evaluation: $M = 3.03$, $t = -5.53.1$, $p < .001$) were significantly below the midpoint. This lends support to the prediction that participants would not consider the agents blameworthy when the vignette clearly states that the agents were not at fault for their inabilities.

Ability ratings were further analysed using a 2 (Vignette: Delivery vs. Evaluation) x 2 (Original of Inability: Original vs. No Self-Induced Inability) between-subjects Anova. As predicted, ability ratings did not differ between Original and No Self-Induced Inability, as neither the main effect of Origin of Inability ($F(1, 712) = 3.29$, $p = .07$) nor the interaction of Origin of Inability and Vignette were significant ($F(1, 712) = 0.01$, $p = .09$) (see Table 3). A second 2 x 2 Anova for blame ratings confirmed the prediction that blame ratings were significantly reduced in the No Self-Induced Inability condition, as demonstrated by the significant main effect of Origin of Inability ($F(1, 712) = 154.75$, $p < .001$, $\eta^2 = .18$). The main effect of Vignette was significant ($F(1, 712) = 46.29$, $p < .001$, $\eta^2 = .06$), and the interaction of Origin of Inability and Vignette was also significant ($F(1, 712) = 26.88$, $p < .001$, $\eta^2 = 0.04$.).

Additionally, the analyses revealed differences between vignettes, confirmed by planned contrasts. In the Original condition, ability ratings were higher for *Evaluation* than for *Delivery*, while there was no significant difference between vignettes for blame ratings. In the No Self-Induced Inability condition, however, both types of ratings were significantly higher for *Evaluation* than for *Delivery*.

These results indicate that making it explicit that the agents did not recklessly or negligently cause their own inabilities (the No Self-Induced Inability condition) reduces people's willingness to blame the agent. This finding supports an alternative interpretation of Turri's data – that people in the original design believed the agent to be at fault for their own inability, and that they blamed the agents for that instead of for not doing as they were supposed to. This, by extension, indicates that participants ascribed derivative, not the required direct, moral responsibility.[12]

[12] These results are in line with some more general observations discussed in Walter Sinnott-Armstrong's chapter in this volume.

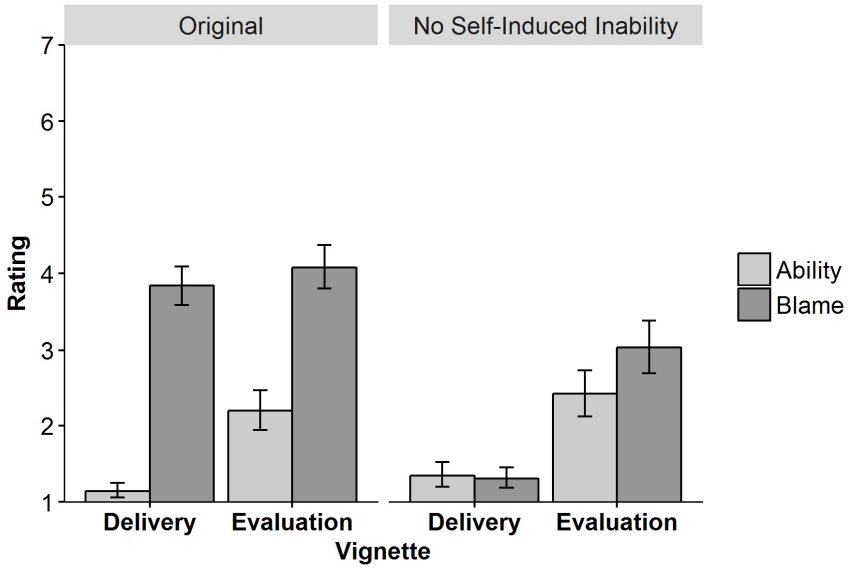

Figure 1: Participant's mean agreement with ability and blame questions in both conditions and vignettes. Error bars indicate 95% confidence intervals.

Likelihood Ratings

In all four conditions, participants judged it rather unlikely that the agent could still do as they were supposed to (Original Delivery: $M = 3.55$, $SD = 12.87$, Original Evaluation $M = 9.42$, $SD = 22.21$; No Self-Induced Inability Delivery: $M = 3.68$, $SD = 12.44$, No-Self Induced Inability Evaluation: $M = 10.97$, $SD = 20.05$). An Anova including the factors vignette and origin of inability revealed that they were higher in the *Evaluation* vignettes compared to *Delivery* – $F(1,712) = 23.57$, $p < .001$, $\eta^2 = 0.03$. These estimates correspond to the generally low (yet higher for *Evaluation*) ability ratings.

Explaining Intuitions in Violation of PAP

In the two Original conditions, 210 out of 304 participants indicated that the agent could not have acted otherwise (agreement to the ability question lower than or equal to 4), but that they were to blame for the consequences of their actions (agreement to the blame question above or equal to 4)[13]. Thus, more than two-thirds of all participants give answers that seem to conflict with PAP. When asked to justify their judgment, 140 participants indicated that the

[13] One might wonder why I chose to include the neutral midpoint. A rating right in the middle between 'strongly disagree' and 'strongly agree' is most likely to express indecisiveness and cannot be interpreted either in favour or disfavour of PAP. I believe that including the neutral midpoint tips the scale in favour of intuitions in violation of PAP and thus makes it easier for Turri to argue in favour of what he calls 'folk compatibilism'. However, I decided to follow Turri in this decision. At the end, the aim of this paper is to examine the validity of Turri's given his premises.

agent was to blame *because* they could have done something to prevent the outcome. In contrast, only a third (70) of all participants who seem to reject PAP (210 out of 304 participants) indicated that the agent was to blame *although* there was nothing they could have done to prevent it.

For *Delivery*, 88% of participants said that the agent was to blame *because* there was something he could have done to prevent the outcome; 45% of participants stated the same for *Evaluation* (see Figure 2). More people agreed that there was something the agent could have done in *Delivery*, as compared to *Evaluation* ($\chi^2 = 49.62$, $p < .001$). Thus, more people gave a response incompatible with PAP for *Evaluation*. Following this specific study, I cannot offer empirical evidence which can explain this difference. However, as previously mentioned, *Delivery* and *Evaluation* differ in a series of potentially important respects, such as familiarity with the situation, relevant background knowledge about brain chemistry, and how external or internal to the agent the inability is. I submit that the most likely explanation is that we all know that when we are running late, there was usually something we could have done to prevent it.

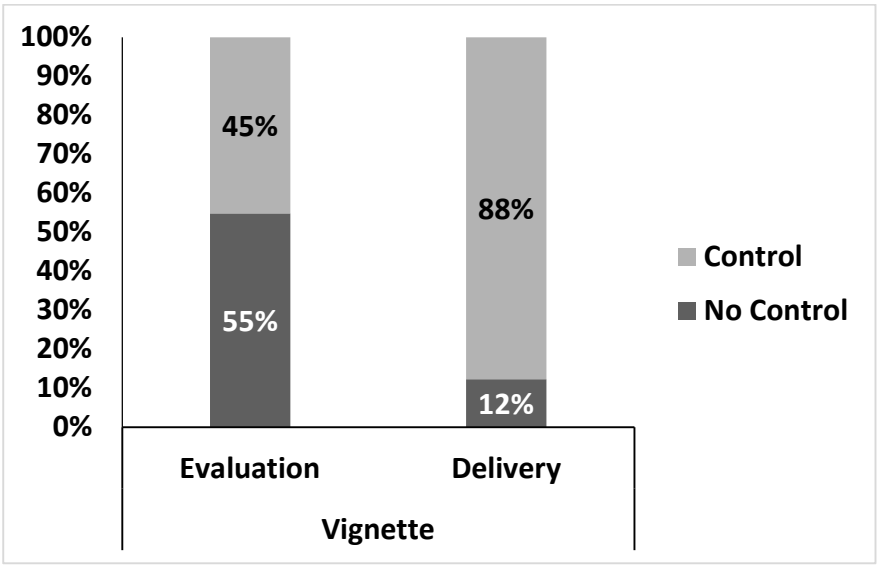

Figure 2: Percentages of people indicating that the agents were to blame although they had no control or because they had control in the Original condition.

To sum up: The Original version of the experiments allowed for the attribution of derivative moral responsibility. The results suggest that the agents are not held directly morally responsible, but that participants hold them responsible by virtue of failing to take adequate precautions to avoid their own inabilities.

Study 1 B

Study 1 B tested the same experimental design as Study 1 A, but for moral responsibility instead of blame ratings.

Participants

710 participants were recruited on the UK-based internet platform *Prolific* (https://www.prolific.ac). All participants were compensated for their participation (0.25 GBP, estimated 7.50 GBP per hour). All participants were native speakers of English and had not previously participated in an experiment using the same vignettes. I excluded 80 participants from the analysis for either failing the attention check, not completing the survey, or finishing the survey in less than 40 seconds. Results are reported for 630 participants ($M = 33.84$, $SD = 10.97$, 61% female, 49% male).

Results

Ability and Moral Responsibility

As in Study 1A, I conducted t-tests against the midpoint of the scale for ability ratings. As predicted, I replicated Turri's results for the Original condition. Participants' ability ratings were significantly below the midpoint for both vignettes (Delivery: $M = 1.47$, $t = -24.99$, $p < .001$; Evaluation: $M = 2.52$, $t = -9.83$, $p < .001$), indicating that they judged the agent unable to do as they were supposed to. Further t-tests against the midpoint for responsibility ratings revealed that responsibility ratings were not significantly different from the midpoint in the case of *Delivery* (Delivery: $M = 4.17$, $t = 1.19$, $p = .2$), while they were significantly above the midpoint in the case of *Evaluation* (Evaluation: $M = 5.04$, $t = 6.98$, $p < .001$) (see Figure 3).

For the No Self-Induced Inability condition, a t-test against the midpoint of the scale (4) revealed that ability ratings (Delivery: $M = 1.33$, $t = -31.85$, $p < .001$; Evaluation: $M = 2.87$, $t = -5.71$, $p < .001$) were significantly below the midpoint. Moral responsibility ratings for *Delivery* were significantly below the midpoint ($M = 2.89$, $t = -6.75$, $p < .001$), while for *Evaluation* they were significantly above the midpoint ($M = 4.43$, $t = 2.37$, $p < .05$)

Ability ratings were further analysed using a 2 (Vignette: Delivery vs. Evaluation) x 2 (Original of Inability: Original vs. No Self-Induced Inability) Anova. Ability ratings did not differ between Original and No Self-Induced Inability, as neither the main effect of Origin of Inability ($F(1, 626) = 0.52$, $p = .47$) nor the interaction of Origin of Inability and Vignette was significant ($F(1, 626) = 0.01$, $p = .08$). I conducted a second 2 x 2 Anova for responsibility ratings. As predicted, there was a significant main effect of Origin of Inability ($F(1, 626) = 34.72$, $p < .001$, $\eta^2 = 0.05$), as in the No Self-Induced Inability condition responsibility

ratings were significantly lower compared to Original. Against my predictions, the manipulation had different strong effects on the two stories, as indicated by the significant two-way interaction ($F(1, 626) = 4.61$, $p < .05$, $\eta^2 = 0.05$). Responsibility ratings decreased significantly below the midpoint for the *Delivery* vignette. In *Evaluation*, responsibility ratings were still significantly above the scale midpoint, indicating that participants still judged the woman to be morally responsible for not meeting her obligation, even though the vignette clearly stated that she was not at fault.

The analyses revealed differences between vignettes, as confirmed by planned contrasts. In the Original condition, ability ratings were again higher for *Evaluation* than for *Delivery*. Moral responsibility ratings differed between vignettes, such that they were higher for *Evaluation* than for *Delivery* across both conditions. In addition, making it clear that the agent was blameless for actions before becoming unable to complete their obligations had a greater effect on moral responsibility ratings for *Delivery* than for *Evaluation*. This lends further support to the claim that the vignettes differ in important theoretical respects.

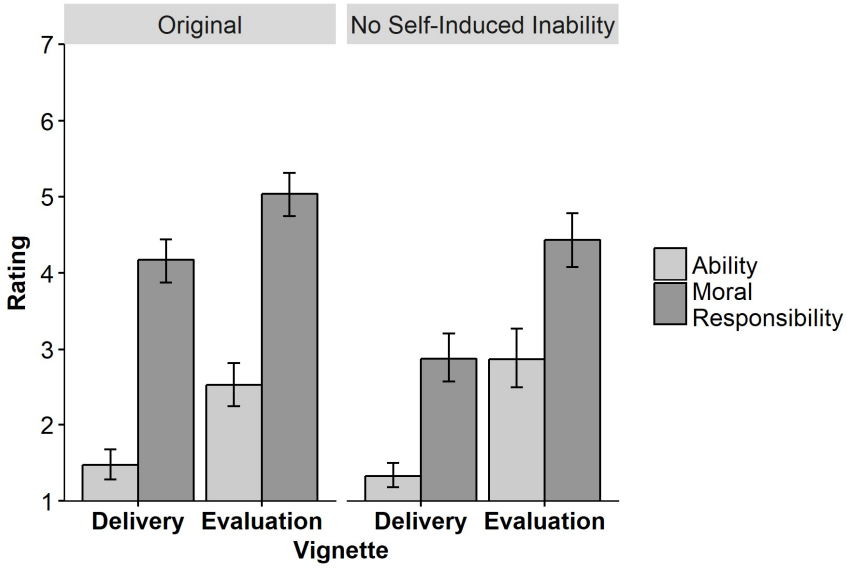

Figure 3: Participants' mean agreement with ability and moral responsibility questions in both conditions and vignettes. Error bars indicate 95% confidence intervals.

Likelihood

In all four conditions, participants judged it rather unlikely that the agent could still do as they were supposed to (Original Delivery: $M = 4.38$, $SD = 11.22$, Original Evaluation: $M = 16.61$, $SD = 27.89$; No Self-Induced Inability Delivery: $M = 6.26$, $SD = 17.43$, No Self-Induced Inability Evaluation: $M = 19.48$, $SD = 25.94$). Confirming the findings of Study 1 A, an Anova which included the factors vignette and origin of inability revealed that estimates were higher for the *Evaluation* vignettes than the *Delivery* ones, $F(1,626) = 52.25$, $p < .001$, $\eta^2 = .08$).

Explaining Intuitions in Violation of PAP

For the two Original conditions, 213 out of 630 participants indicated that the agent could not have acted otherwise (agreement to the ability question lower than or equal to 4), but that they were to blame for the consequences of their actions (agreement to the blame question above or equal to 4). When asked to justify their judgments in violation of PAP, 105 participants indicated that the agent was to blame *because* they could have done something to prevent the outcome. 108 participants who seemed to reject PAP indicated that the agents were to blame *although* there was nothing they could have done to prevent it.

This time, answers in the Original condition did not differ between vignettes ($\chi^2 = 0.04$, $p = 0.83$, *n.s.*). Both in *Evaluation* and in *Delivery*, participants' choices were distributed equally between the claim that the agent was to blame *because* there was something they could have done to prevent the outcome and the claim that they were to blame *although* there was nothing they could have done (see Figure 4).

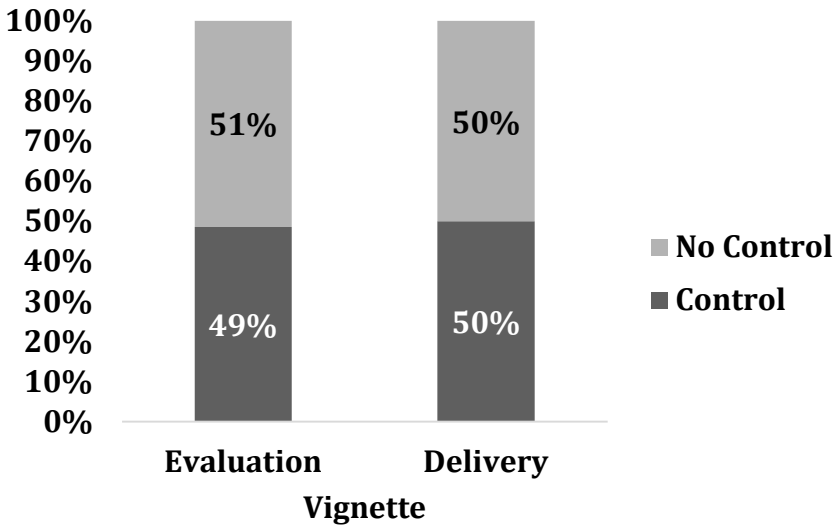

Figure 4: Percentages of people indicating that the agents were to blame although they had no control or because they had control in the Original condition.

Discussion

Studies 1 A and B challenge the reliability of Turri's compatibilist conclusion in two ways. First, the results lend support to the view that participants did not ascribe direct moral responsibility to the agent for their actions, but instead ascribed derivative moral responsibility. Excluding some of the potentially most obvious things that they should have done differently, such as leaving earlier or paying better attention to the traffic news,

significantly reduced moral responsibility attribution. This effect occurred for both moral responsibility measures. In contrast, manipulating the origin of the agents' inabilities did not affect ability ratings. Adding information that the agents did not cause their own inabilities therefore only and directly affected the agents' moral status. These results are compatible with the idea that the moral responsibility that was measured was derivative, not direct, moral responsibility, and they put severe pressure on the validity of Turri's original interpretation in favour of folk compatibilism.

Second, the results also indicate more directly that participants did not buy into the necessary premise to test the acceptance of PAP, namely the agent's inability to act otherwise. More specifically, it seems that a large portion of the responses which seem to stand in conflict with PAP can be explained by participants' beliefs that there was something the agent could have done to avoid delivering the package late or writing a bad evaluation. As these results demonstrate, asking participants to explain their seemingly PAP-violating judgments can serve as a control for whether a judgment is actually in violation of PAP or not.

5 Study 2

The experimental manipulation described in Studies 1 A and B could raise potential concerns. Between the original and the manipulated conditions, vignettes do not differ only in the extent to which the agent could be considered blameworthy for something before the inability manifested. The manipulated vignettes are also, as a necessary consequence of the manipulation, significantly longer, and entail additional factors such as the police announcing the closure of the route or the stroke from which the woman suffers. As an alternative to my suggestion, one might believe that what really explains the effects between Original and No Self-Induced Inability vignettes is not the agents' recklessness, but the length of the vignettes (which has now doubled) or the introduction of additional agents or other variables.

Thus, in this second experiment, I directly tested this possibility. To reach a comparable length, I extended Turri's original *Delivery* vignette by adding irrelevant information with respect to the agent's blameworthiness. I predicted that making the Original conditions longer would not result in blame being significantly reduced. The experimental design and all predictions and analyses were again pre-registered with the open science framework (*https://osf.io/82ems/?view_only=c229aede0d19439b83bcddb31ad938da*).[14]

[14] In this experiment, I only used the *Delivery* vignette and did not also test the *Evaluation* vignette. There are several reasons for this decision. First, as mentioned in Section 3, it is unclear whether participants even understand the connection between brains, chemistry, and an agent's behaviour. The *Delivery* vignette seems sufficiently intelligible. Second, the *Evaluation* vignette further leaves room for several interpretations according

Methods

Participants

343 participants were recruited on the UK-based internet platform *Prolific* (https://www.prolific.ac). All participants were compensated for their participation (0.25 GBP, estimated 7.50 GBP per hour). All participants were native speakers of English and had not previously participated in an experiment using the same vignettes. I excluded 46 participants from the analysis for either failing the attention check, not completing the survey, or finishing the survey in less than 40 seconds. Results are reported for 297 participants ($M = 36.10$, $SD = 10.95$, 66% female, 34% male).

Design and Procedure

I tested three between-subjects conditions, namely Original vs. No Self-Induced Inability vs. Original Long for the *Delivery* vignette. The Original and the No Self-Induced Inability conditions were identical to those used in Studies 1 A and B. In Original Long, I took the Original version and added irrelevant information in those places in which the No Self-Induced Inability condition contains information about the agent not being blameworthy for his inability.

The new modified Original Long vignette now reads:

> A man promised to deliver a package by 4 pm. Before he leaves, he checks the results of yesterday night's football games and how his favourite player performed.
>
> He just got on the freeway, when he hears on the radio that the police announce an open day at the local police station next week Sunday. Young people interested in becoming a police officer can visit and ask questions about the job and the entry conditions. There will also be music and a bouncy castle for children.
>
> Given current traffic conditions, it is physically impossible that the man can deliver the package by 4 pm. As a matter of physics, it is literally impossible that he can make it by 4 pm. He will arrive late.

After reading one of the three vignettes, participants answered the ability, responsibility, and likelihood questions from Study 1. Unlike in Studies 1 A and B, participants did not answer the time estimate question for *Delivery*, and neither were participants with seemingly PAP-violating intuitions asked to explain their judgments.

---

to which the agent does have alternative possibilities when performing the action in question. One might think, for instance, that the negative evaluation is conditional on the woman writing the evaluation right now. However, so participants might reason, she does not have to write the evaluation *now*. The vignette only specifies that the woman is going to write a negative evaluation, as opposed to a positive one; it does not specify that she has to write the evaluation now, as opposed to writing it later. What is worse, it remains unclear whether the woman is aware of her own neurological status. If participants think that she might be and that she could write the evaluation at a later point, it is clear that the woman is directly responsible for the bad evaluation and should not have written it in the first place.

Results and Discussion

As Figure 5 shows, making the vignettes longer did not reduce blame ratings. In the No Self-Induced Inability Condition, blame was significantly lower compared to Original (M = 3.58, SD = 2.0 for Original, M = 1.49, SD =1.18 for Not Self-Induced, t = 8.9, p < .001, r = .58) or to Original Long (M = 4.68, SD = 1.79, t = 14.98, r = .75). In line with my explanation for the effect found in Study 1, making the original vignettes longer (Original Long condition) did not by itself decrease blame ratings. Only when I provided blame-relevant information (No Self-Induced Inability) did blame ratings drop. These results provide evidence that the manipulation was successful, and that making it explicit that the agent is blameless for their inability reduces blame ratings.

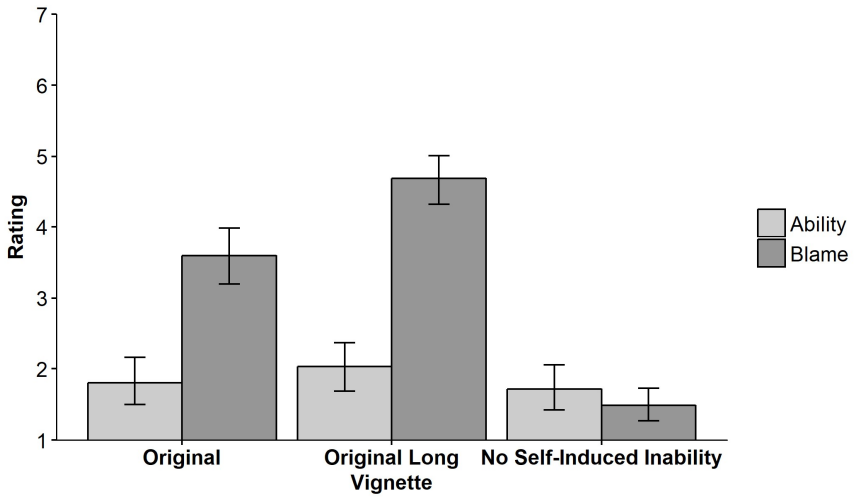

Figure 5: Participants' mean agreement with ability and moral responsibility questions in all three conditions. Error bars indicate 95% confidence intervals.

Against my initial prediction, blame ratings for Original Long were not equally as high as blame ratings for Original, but were significantly higher (t = 4.0, p < .001, r = .28). However, I believe that this effect should not be overstated, as the effect might have been driven largely by participants assuming a connection between checking football results and not delivering the package in time. A reasonable assumption is that the delivery man left too late *because* he checked football results before leaving. However, the fact that introducing (arguably) irrelevant information to the Original Delivery vignette *increased* blame ascriptions instead of *decreasing* them only strengthens my argument that it is actually the content of the additional information that matters.

6 General Discussion

Do the folk accept or reject the Principle of Alternative Possibilities? Experimental philosophers have attempted to provide an empirical answer to this question and, thereby, to inform the traditional, non-empirical debate. I argue that in order to provide evidence for the folk rejecting PAP, three things must be demonstrated:

1. Participants say that the agent could not have acted otherwise than performing X at t.
2. Those participants who believe that the agent could not have acted otherwise still ascribe moral responsibility for X at t.
3. Participants ascribe direct moral responsibility.

Discussing and examining Turri (2017a), I argue that his experiments indeed show 1 and 2. The goal of this paper was to show that 1 and 2 alone do not suffice to make the point that the folk reject PAP, let alone are compatibilists. Rather, one also has to show that 3 is the case. I raised the concern that condition 3 is has not been justified, and that people in fact ascribe derivative (rather than direct) moral responsibility. I conducted three experiments to put this hypothesis to the test. Supporting my suspicion that participants tend to ascribe derivative moral responsibility, I demonstrated that revising the original vignettes by adding information specifying that the agent did not negligently or intentionally cause their own inability (something that might provide the grounds for derivative moral responsibility) significantly reduced moral responsibility ratings. Adding this information did not alter participants' ability or likelihood ratings, but only had an effect on moral responsibility ratings. The same effects occurred for blame as an alternative measure for moral responsibility. Additional questions asking for an explanation of judgments which seem to reject PAP revealed that people believed that there was something the agent could have done to prevent their own inability. These explanations strongly support the view that the kind of moral responsibility that participants ascribed was derived moral responsibility, not the direct moral responsibility we should require.

The effect found in this paper has far-reaching implications, as it points to a general methodological issue with many studies in experimental philosophy. Variations of the vignettes discussed in this paper and in Turri (2017) feature also in other publications that are often cited in the experimental literature, such as Turri (2017b, 10 citations[15]), Buckwalter and Turri (2015, cited 50 times), Henne et al. (2016, cited 27 times), and Chituc et al. (2016, cited 55 times). While many authors have been critical of this evidence, it continues to have a

[15] All citations are based on GoogleScholar and were last checked on 23 March 2021.

significant impact on the philosophical debate (Kissinger-Knox et al., 2018; Kurthy et al., 2017; Streumer, 2003; Willemsen & Wiegmann, 2017). In addition, most experimental studies rely on experimental stimuli that resemble the two vignettes described in this study, in that an agent is described as determined in conducting and acting in a specific situation (Buckwalter, 2017; Miller & Feltz, 2011; Murray & Lombrozo, 2017; Willemsen, 2018, 2020; Woolfolk et al., 2006).

As no experiment that I am aware of has tested for the possibility that participants ascribe derivative moral responsibility instead of direct moral responsibility[16], we should be careful when drawing any philosophical conclusions from this evidence until follow-up studies confirm that participants ascribe direct moral responsibility. While the experimental stimuli are explicitly reported in these papers, they are often omitted in summary articles (which solely focus on the results of these studies) on the advances in experimental philosophy of compatibilism and PAP (Semler & Henne, 2019). Thus, experimental stimuli that are prone to triggering the attribution of derivative instead of direct moral responsibility are repeatedly used in experimental studies and their results are summarised in overview articles, hindering critical reflection on the stimuli and test queries.

7 Funding Information and Acknowledgement

This research was funded by the Swiss National Science Foundation (SNSF), grant number PCEFP1_181082. I would like to express my gratitude to Sabrina Coninx, Neele Engelmann, Lena Kaestner, Beate Krickel, Matthew Lindauer, Judith Martens, Thomas Nadelhoffer, Kevin Reuter, Simon Stephan, and Alex Wiegmann for providing invaluable feedback on earlier versions of this paper.

[16] It might be argued that all studies on omissions, negligence, and recklessness necessarily deal with derivative rather than direct moral responsibility. In a case of negligence, to say that the agent was negligently responsible is to say that the agent is derivatively morally responsible for the outcome in virtue of failing to anticipate the risks involved in acted the way they did and failing to take adequate precautions. I fully agree that in these studies, derivative moral responsibility is likely to be investigated. The point I wish to make is that the question of whether direct of derivative responsibility is ascribed in these cases has not been empirically addressed.

8 References

Buckwalter, W. (2017). Ability, responsibility, and global justice. *Journal of Indian Council of Philosophical Research*, *34*(3), 577–590. https://doi.org/10.1007/s40961-017-0120-z

Buckwalter, W., & Turri, J. (2015). Inability and obligation in moral judgment. *PLOS ONE*, *10*(8), e0136589. https://doi.org/10.1371/journal.pone.0136589

Chituc, V., Henne, P., Sinnott-Armstrong, W., & De Brigard, F. (2016). Blame, not ability, impacts moral "ought" judgments for impossible actions: Toward an empirical refutation of "ought" implies "can". *Cognition*, *150*, 20–25. https://doi.org/10.1016/j.cognition.2016.01.013

Fischer, J. M., & Ravizza, M. (1998). *Responsibility and control: A theory of moral responsibility* (1. Edition). Cambridge University Press. https://doi.org/10.1017/CBO9780511814594

Frankfurt, H. G. (1969). Alternate possibilities and moral responsibility. *The Journal of Philosophy*, *66*(23), 829–839. https://doi.org/10.2307/2023833

Ginet, C. (2000). The epistemic Requirements for moral responsibility. *Nous*, *34*(s14), 267–277. https://doi.org/10.1111/0029-4624.34.s14.14

Henne, P., Chituc, V., De Brigard, F., & Sinnott-Armstrong, W. (2016). An empirical refutation of 'ought' implies 'can'. *Analysis*, *76*(3), 283–290. https://doi.org/10.1093/analys/anw041

Huber, L., Reuter, K.; Cacchione (forthcoming). Children and Adults Don't Think They Are Free:A Skeptical Look at Agent Causationism. In Pascale Willemsen & Alex Wiegmann (edit.), *Advances in Experimental Philosophy of Causation*. London: Bloomsbury Press.

Kane, R. (1999). Responsibility, luck, and chance: Reflections on free will and indeterminism. *The Journal of Philosophy*, *96*(5), 217–240. https://doi.org/10.2307/2564666

Khoury, A.C. (2012). Responsibility, Tracing, and Consequences. *Canadian Journal of Philosophy*, 42(3–4), 187–207. doi:10.1080/00455091.2012.10716774

King, M. (2014). Traction without Tracing: A (Partial) Solution for Control-Based Accounts of Moral Responsibility. *European Journal of Philosophy*, 22(3), 463–482. doi:10.1111/j.1468-0378.2011.00502.x

Kissinger-Knox, A., Aragon, P., & Mizrahi, M. (2018). "Ought implies can," Framing effects, and "empirical refutations". *Philosophia*, *46*(1), 165–182. https://doi.org/10.1007/s11406-017-9907-z

Kurthy, M., Lawford-Smith, H., & Sousa, P. (2017). Does ought imply can? *PLOS ONE*, *12*(4), e0175206. https://doi.org/10.1371/journal.pone.0175206

Levy, N. (2017). The good, the bad, and the blameworthy. *Journal of Ethics and Social Philosophy*, *1*(2), 1–16. https://doi.org/10.26556/jesp.v1i2.6

Lycan, W. G. (2003). Free will and the burden of proof. *Royal Institute of Philosophy Supplement*, *53*, 107–122. https://doi.org/10.1017/S1358246100008298

Matheson, B. (2019). Towards a structural ownership condition on moral responsibility. *Canadian Journal of Philosophy*, *49*(4), 458–480. https://doi.org/10.1080/00455091.2018.1480853

Mele, A. R. (2020). Direct Versus Indirect: Control, Moral Responsibility, and Free Action. *Philosophy and Phenomenological Research*, phpr.12680. https://doi.org/10.1111/phpr.12680

Miller, J. S., & Feltz, A. (2011). Frankfurt and the folk: An experimental investigation of Frankfurt-style cases. *Consciousness and Cognition*, *20*(2), 401–414. https://doi.org/10.1016/j.concog.2010.10.015

Murray, D., & Lombrozo, T. (2017). Effects of manipulation on attributions of causation, free will, and moral responsibility. *Cognitive Science*, *41*(2), 447–481. https://doi.org/10.1111/cogs.12338

Nahmias, E., Morris, S., Nadelhoffer, T., & Turner, J. (2005). Surveying freedom: folk intuitions about free will and moral responsibility. *Philosophical Psychology*, *18*(5), 561–584. https://doi.org/10.1080/09515080500264180

Nelkin, D. K.; Rickless, S. C. (2017). Moral Responsibility for Unwitting Omissions: A New Tracing View. In Nelkin and Rickless (edit), *The Ethics and Law of Omissions*, Oxford: Oxford University Press.: 106–130.

Robichaud, P., & Wieland, J. W. (2017). *Responsibility: The epistemic condition*. Oxford University Press.

Rosen, G. (2003). Culpability and Ignorance. *Proceedings of the Aristotelian Society*, 103(1), 61–84. doi:10.1111/j.0066-7372.2003.00064.x

Rosen, G. (2004). Skepticism about moral responsibility. *Philosophical Perspectives*, *18*(1), 295–313. https://doi.org/10.1111/j.1520-8583.2004.00030.x

Rudy-Hiller, F. (2018). The Epistemic condition for moral responsibility. In *The Stanford Encyclopedia of Philosophy*. https://plato.stanford.edu/archives/fall2018/entries/moral-responsibility-epistemic/

Shabo, S. (2015). More Trouble with Tracing. *Erkenntnis*, 80(5), 987–1011. doi:10.1007/s10670-014-9693-y

Semler, J., & Henne, P. (2019). Recent experimental work on "ought" implies "can". *Philosophy Compass*, *14*(9). https://doi.org/10.1111/phc3.12619

Sommers, T. (2010). Experimental philosophy and free will. *Philosophy Compass*, *5*(2), 199–212. https://doi.org/10.1111/j.1747-9991.2009.00273.x

Strawson, G. (1994). The impossibility of moral responsibility. *Philosophical Studies*, *75*(1–2), 5–24. https://doi.org/10.1007/BF00989879

Streumer, B. (2003). Does "ought" conversationally implicate "can"? *European Journal of Philosophy*, *11*(2), 219–228. https://doi.org/10.1111/1468-0378.00184

Timpe, K. (2011). Tracing and the Epistemic Condition on Moral Responsibility. *The Modern Schoolman*, 88(1-2): 5–28. doi:10.5840/schoolman2011881/22

Turri, J. (2017a). Compatibilism can be natural. *Consciousness and Cognition*, *51*, 68–81. https://doi.org/10.1016/j.concog.2017.01.018

Turri, J. (2017b). How "ought" exceeds but implies "can": Description and encouragement in moral judgment. *Cognition*, *168*, 267–275. https://doi.org/10.1016/j.cognition.2017.07.008

Van Inwagen, P. (1975). The incompatibility of free will and determinism. *Philosophical Studies*, *27*(3), 185–199. https://doi.org/10.1007/BF01624156

Vargas, M. (2006). On the importance of history for responsible agency. *Philosophical Studies*, *127*(3), 351–382. https://doi.org/10.1007/s11098-004-7819-9

Willemsen, P. (2018). Omissions and expectations: A new approach to the things we failed to do. *Synthese*, *195*(4), 1587–1614. https://doi.org/10.1007/s11229-016-1284-9

Willemsen, P. (2020). The relevance of alternate possibilities for moral responsibility for actions and omissions. In T. Lombrozo, J. Knobe, & S. Nichols (Hrsg.), *Oxford studies in experimental philosophy* (Bd. 3). Oxford University Press.

Willemsen, P., & Wiegmann, A. (2017). *I must although I can't!? Suggestions for a two-level-theory of "ought implies can"*. https://doi.org/10.31234/osf.io/hyq9u.

Woolfolk, R. L., Doris, J. M., & Darley, J. M. (2006). Identification, situational constraint, and social cognition: Studies in the attribution of moral responsibility. *Cognition*, *100*(2), 283–301. https://doi.org/10.1016/j.cognition.2005.05.002

Zimmerman, M. J. (1997). Moral responsibility and ignorance. *Ethics*, *107*(3), 410–426. https://doi.org/10.1086/233742